Entwicklung und die Überprüfung der psychometrischen Eigenschaften eines standardisierten Verfahrens zur Messung von Empowerment im Prozess der psychiatrischen Behandlung von Patienten mit affektiven (ICD-10 F30-F39) und schizophrenen (ICD-10 F20-F29) Störungen.

Förderkennz. 01GX0743

Projektleitung: PD Dr. Reinhold Kilian

Abschlussbericht

Anhang III: Methodendarstellung IRT Analysen

Dr. Herbert Matschinger

Institut für Sozialmedizin, Arbeitsmedizin und Public Health Universität Leipzig, Medizinische Fakultät

Korrespondenzandresse:

Institut für Sozialmedizin, Arbeitsmedizin und Public Health Universität Leipzig,
Medizinische Fakultät http://www.uni-leipzig.de/~sasm/
Philipp-Rosenthal-Straße 55, 04103 Leipzig
Tel.: +49 (0)341/9724532, Fax: +49 (0)341/9724569
e-mail: Herbert.Matschinger@medizin.uni-leipzig.de

Methodische Zugänge

Im folgenden sollen die zur Dimensionsbestimmung und Variablenselektion notwendigen Begriffe und Methoden definiert und in der gebotenen Kürze dargestellt werden. Die hier beschriebenen Prozeduren und Vorgehensweisen gelten prinzipiell für die Pilot- und für die Feldstudie. Allerdings werden für die Pilotstudie sowohl exploratorische, wie konfirmatorische Analysen durchgeführt. Für die Feldstudie sind nur noch Zugänge vorgesehen, welche die dimensionale Struktur des Instrumentes nicht mehr verändern, sondern innerhalb derselben nach einer sparsameren Repräsentation suchen. Ziel ist es, für jede Dimension die Indikatoren so auszuwählen, dass strukturelle Invarianz unter definierten Bedingungen gegeben ist. Strukturelle Invarianz bezeichnet die Tatsache gleicher Messeigenschaften eines Instruments unter definierten Bedingungen.

Für die Pilotstudie werden auch exploratorische Ansätze zur Analyse der Dimensionsstruktur im Sinne der „construct maps“ (Wilson, 2005). Im ersten Schritt werden exploratorische Faktoranalysen (WLMSV) mit GEOMIN (oblique) Rotation durchgeführt (Sass & Schmitt, 2010). Diese Analysen sind exploratorisch. Der konfirmatorische Zugang für kategoriale Variable geschieht im Rahmen der IRT Analysen. Die Vorteile dieser Methodik ist vielerorts diskutiert; stellvertretend sei hier nur eine Arbeit aufgeführt (Reeve, Hays, Chih-Hung et al., 2007)

Bei der Beurteilung der Messeigenschaften ist zu beachten, dass sich Items immer aus jeweils 2 hierarchisch geordneten Stimuli zusammensetzen. Die folgende Definition macht dies deutlich:

“A test item in an examination of mental attributes is a unit of measurement with a **stimulus and a prescriptive form of answering**, and, it is intended to yield a response from an examinee from which perfomance in some psychological **construct** (such as ability, predisposition, or trait) may be inferred“ (Reckase, 2009)

Daraus lassen sich jene notwendigen Analysen herleiten, die im Folgenden zur Charakterisierung der Subdimensionen des Konzeptes „Empowerment“ herangezogen werden. Da für alle Items ausschließlich 5 benannte Antwortkategorien ordinaler Natur eingesetzt werden, wird ausschließlich das PCM (Partial Credit Modell) (Masters, 1982; Masters, 1988) herangezogen, da sich ein Ratingscale und noch stärker restringierte Modelle

(Andrich, 1978; Andrich, 1982; Wright & Masters, 1982) in jedem Fall verbieten. Das Modell ist wie folgt definiert:

$$P(y_{ip} = j \mid \theta_p, \delta_{il}) = \frac{\exp \sum_{l=0}^{j} (\theta_p - \delta_{il})}{\sum_{k=0}^{m_i} \exp \sum_{l=0}^{k} (\theta_p - \delta_{il})} \qquad \sum_{l=0}^{0} (\theta_p - \delta_{il}) = 0$$

bzw.

$$P(y_{ip} = j \mid \theta_p) = \frac{\exp(\nu_{ijp})}{\sum_{k=0}^{m_i} \exp(\nu_{ijp})} \qquad j = 0,1,.....m_i$$

Das Partial Credit Modell spezifiziert die Wahrscheinlichkeit der Kategorie *j* bei Item *i* für die Person *p* als eine Funktion der Fähigkeit θ_p und der Schwellenparameter δ_{il}. Dies ist ein Spezialfall des „adjacent category logit model“ (Agresti, 1989; Agresti, 2002)

$$\ln \frac{P(y_{ip} = j \mid \theta_p)}{P(y_{ip} = j-1 \mid \theta_p)} = \theta_p - \delta_{ij}$$

In jedem Fall wird angenommen, dass ein bestimmtes Modell unter alle Bedingungen gilt, die Modellparameter, gegeben θ_p , bis auf eine Konstante gleich sind. Gilt dies nicht, so liegt DIF bzw. im mehrkategoriellen Fall ein unterschiedliches Funktionieren der Reaktionskategorien vor.

Diese wenig wünschenswerte Heterogenität kann sich aus sehr unterschiedlichen Quellen speisen, wobei zunächst die Eigenschaften der Befragten und die Eigenschaften der Stimuli als Hauptursachen zu unterscheiden sind. Zu den Eigenschaften der Befragten zählt dabei auch der jeweils individuell gültige Wert auf der gemessenen (latenten) Dimension. Ein wesentliches Charakteristikum der Strukturellen Invarianz ist auch das Fehlen von „differential item functioning“ (DIF). Dabei ist bei mehrkategoriellen Fragen aber auch das

differentielle Funktionieren der Kategorien zu untersuchen. Ganz allgemein lässt sich aber sagen, dass UCI (Unobserved Conditional Invariance) gegeben ist, wenn gilt:

$$P(y_{ip}|\theta=\theta_p, V=v) = P(y_{ip}|\theta=\theta_p)$$

Es sei y_{ip} die Messung für θ und V eine beliebige Charakteristik der Befragten. Nur wenn die obige Relation für alle V und θ gilt ist die Messung der latenten Dimension unabhängig von allen Werten v. Mit anderen Worten: y ist eine unverzerrte Messung der latenten Dimension dann und nur dann, wenn UCI gegeben ist.

$$P(y_{ip}| V=v) \neq P(y_{ip})$$

Die Ungleichheit der beiden Wahrscheinlichkeiten allein ist noch KEIN Indikator für eine verzerrte Messung, weil Befragte mit unterschiedlichen Eigenschaften v durchaus sehr verschiedene Werte auf θ aufweisen können, was wiederum zu unterschiedlichen Wahrscheinlichkeiten für y unter der Bedingung θ führen kann (Engelhard, Jr., 2009; Eun & Myeongsun, 2011; Teresi & Fleishman, 2007; Zumbo, 2007).

Um die beschriebene Heterogenität (DIF) zu modellieren, verwenden wir ein etwas andere Darstellung des PCM (De Boeck, 2008). Das Modell kann als Spezialfall eine verallgemeinert IRT Modells gesehen werden (Jannsen, Scheper & Peres, 2004; Rijmen, Tuerlinckx, De Boeck et al., 2003; Wilson & De Boeck, 2004) und lässt sich daher auch als „generalized linear latent“ Modell notieren (Rabe-Hesketh, Skrondal & Pickels, 2004; Skrondal & Rabe-Hesketh, 2004):

$$\nu_p = X_p\beta_i + \theta_p Z_p\lambda \qquad \nu \cong g\left(E[y \mid x,\theta]\right)$$

Als Verknüpfungsfunktion g wird hier die logistische Funktion angenommen. Für eine 3-kategorielle Variable lässt sich der Numerator v wie folgt notieren:

$$\text{If } j = 0,\ v_{i0p} = 0$$
$$\text{If } j = 1,\ v_{i1p} = 0 + (\theta_p - \delta_{i1})$$
$$\text{If } j = 2,\ v_{i2p} = 0 + (\theta_p - \delta_{i1}) + (\theta_p - \delta_{i2}) = 2\,\theta_p - \delta_{i1} - \delta_{i2}$$

Für 2 Items mit je 3 Kategorien (0,1 und 2) stellt sich das konditionale Logit Modelle wie folgt dar (Zheng & Rabe-Hesketh, 2007: 316f):

$$\underbrace{\begin{bmatrix} \nu_{10p} \\ \nu_{11p} \\ \nu_{12p} \\ \nu_{20p} \\ \nu_{21p} \\ \nu_{22p} \end{bmatrix}}_{\nu_p} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & -1 \end{bmatrix}}_{X_p} \begin{bmatrix} \delta_{11} \\ \delta_{12} \\ \delta_{21} \\ \delta_{22} \end{bmatrix} + \theta_p \underbrace{\begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix}}_{Z_p}$$

Jeder Schwellenwert ist durch eine Variable im Modell repräsentiert. Um Schwierigkeitsparameter zu erhalten sind diese Variablen mit –1 notiert. Das Modell kann durch weitere, unabhängige Variablen erweitert werden, wobei diese sowohl auf der Item- wie auch auf der Befragtenebene variieren können (De Boeck & Wilson, 2004; Meulders & Xie, 2004; Wilson & De Boeck, 2004). Durch geeignete Restriktionen dieser Effekte auf die Schwellenwerte – z.B. durch Gleichheitsrestriktion zwischen den Items – kann dann DIF näher untersucht werden. Zu diesem Zweck werden für jede Schwellenwertvariable Interaktionen mit den vermuteten DIF-Variablen gebildet und damit die entsprechenden Effekte modelliert. Item***un***spezifische Effekte auf die Schwellenwerte liegen vor, wenn die Gleichheitsrestriktion der Effekte für alle Items statistisch akzeptabel erscheint. Werden alle Effekte auf die Schwellenwerte gleichgesetzt, so ist der Effekt einer Befragtencharakteristik durch einen einzigen Parameter repräsentiert. Andererseits ist es möglich, die Effekte der personenspezifischen Variablen auf den Personenparameter θ zu schätzen. Dieses Modell ist formal mit dem zuvor genannten identisch, da es die Invarianz der Effekte auf **alle** Schwellenwerte zur Voraussetzung hat.

Zur Beurteilung jeder der 5 Dimensionen werden jeweils 4 Modelle mit einem Prädiktor für den differentielle Effekte auf die Schwellenwerte vermutet werden, gerechnet. Diese sind:

1. PCM ohne jeden Prädiktor

2. PCM mit Effekt auf die Personenparameter θ

3. PCM mit schwellenwertspezifischem und itemspezifischem Effekt

4. PCM mit für alle Items gleichen, schwellenwertspezifischem Effekt

5. PCM mit für alle Items identischem schwellenwertspezifischem Effekt

Das Modell 5, für das ***alle*** Effekte auf Gleichheit restringiert werden, muss nicht geschätzt werden, da es formal dem Modell 2 entspricht. Die genannten Modell sind hierarchisch geordnet, sodass ein LR-Test möglich ist. Ziel ist hier gleichzeitig den Effekt relevanter Variabler auf die Dimensionen von „Empowerment" und mögliche itemspezifische Effekte zu modellieren. Von Bedeutung ist hier auch die etwas weniger strenge Restriktion auf Schwellenspezifische Effekt, die aber zwischen den Items nicht variieren. Im Anhang finden sich die vollständigen Ausdrucke für alle berechneten Modelle. Alle Berechnungen erfolgten mit dem Modul gllamm (Rabe-Hesketh, Skrondal & Pickles, 2004) für STATA (StatCorp., 2009)

Zur Beurteilung der Brauchbarkeit eines Items werden die Kennwerte „infit" und „outfit" herangezogen. Infit und outfit sind 2 kontrovers diskutierte Beurteilungskriterien für die manifesten Indikatoren einer (eindimensionalen) latenten Variable. Prinzipiell sind sie ein Maß für die „Zufälligkeit" bzw. „Determiniertheit" eines Items mit Rücksicht auf das der latenten Variable zu Grunde liegende Modell. Werte größer 1 kennzeichnen Items mit allzu großer Variationsbreite; die antworten sind durch das Modell nicht ausreichend genau vorhersagbar (*Underfit*). Werte kleiner 1 kennzeichnen hingegen Items mit allzu strenger (deterministischer) Antwortstruktur. Die Variationsbreite ist zu klein, d.h. die Struktur ist zu deterministisch (Das Guttman pattern wird als mit einem probabilistischen Modell unverträglich, weil als zu rigide, betrachtet (Bond & Fox, 2007). Letzteres wird auch „*overfit*" genannt. Die Begriffe *„zu klein"* und *„zu groß"* unterliegen naturgemäß einer inferenzstatistischen Beurteilung, die ebenfalls kontrovers diskutiert wird, da dies, wie immer, ein Problem der Stichprobengröße (sowohl der Items, wie auch der Personen !) ist. Die genannten Statistiken haben einen Erwartungswert von 1 und eine Spannweite von 0 bis $+\infty$. Folgendes Zitat ist hilfreich und bezieht sich auf (Wright & Masters, 1982) http://www.rasch.org/rmt/rmt34e.htm

> Page 100 of Rating Scale Analysis (Wright and Masters 1982), which summarizes the calculation of outfit and infit statistics, is reprinted opposite. Outfit is based on a sum of squared standardized residuals. Standardized residuals are modeled to approximate a unit normal distribution. Their sum of squares approximates a χ^2 distribution. Dividing this sum by its degrees of freedom yields a mean-square value, OUTFIT MEANSQ,

with expectation 1.0 and range 0 to infinity. Values larger than 1.0 indicate unmodeled noise. Values are on a ratio scale, so that 1.2 indicates 20% excess noise. Values less than 1.0 indicate a lack of stochasticity. A Wilson-Hilferty transformation standardizes the mean-square into its OUTFIT ZSTD value. This approximates a unit-normal distribution. Infit is an information-weighted form of outfit. The weighting reduces the influence of less informative, low variance, off-target responses. It is also computed in INFIT MEANSQ and INFIT ZSTD forms. Computation of OUTFIT and INFIT Statistics. Wright BD, Masters GN. … Rasch Measurement Transactions, 1990, 3:4 p.84-5

Bei (Linacre & Wright, 1994) werden diese Kennwerte wie folgt bewertet:

Interpretation of parameter-level mean-square fit statistics:

>2.0	Distorts or degrades the measurement system
1.5 - 2.0	Unproductive for construction of measurement, but not degrading
0.5 - 1.5	Productive for measurement
<0.5	Less productive for measurement, but not degrading. **May produce misleadingly good reliabilities and separations** (Fettdruck durch den Author)

Es bleibt fraglich, ob im Sinne eines probabilistischen Modells zu rigide Items tatsächlich als brauchbare Indikatoren verworfen werden sollen. Unvorhersehbare Reaktionsmuster, also Items mit Fit-Indizes weit größer als 1 („*underfit*“) sollten in jedem Falle ausgeschieden werden, besonders und gerade weil sie als Indikatoren einer bestimmten Subdimension vorgesehen waren und diese offensichtlich zur Erklärung der Reaktionsmuster nur wenig beiträgt. Die Indizes setzen ja die theoretisch gültige Zuordnung voraus. Items mit „*overfit*“ trägt nur wenig zur Messung bei, widerspricht aber nicht den Modellannahmen Die Berechnung der Modelle und der Personenparameter zur Schätzung des Modellfits geschieht mit dem R-Paket eRm (Hatzinger & Mair, 2007; Mair & Hatzinger, 2007). Die Beurteilung der Schwellenwerte beruht auf ihrer Ordnung auf der latenten Dimension, wobei die graphische Darstellung sowohl mit WINMIRA, als auch mit der Funktion plotPImap in R durchgeführt werden kann. Letztere zeichnet die Verteilung der Personenparameter *und* die Spannweite der Schwellenwerte direkt untereinander, was die Interpretation sehr erleichtert.

Da wir annehmen müssen, dass das jeweilige Modell nicht für die gesamte Stichprobe gültig ist, wird dieses sowohl in einer, wie auch in 2 latenten Klassen geschätzt (Rost, 1990; Rost, 1991; Rost, 2004; Rost & von Davier, 1995; von Davier & Rost, 1995). Diskrete Mischverteilungs- IRT Modelle nehmen an, dass die Daten aus einer – nicht beobachtbaren (!) - Mischung von *Populationen* gezogen sind. Damit wird angenommen, dass Modellabweichungen nicht unbedingt „Messfehler“ sind, sondern eben Populationen

existieren für die eine Messung nicht *gültig* ist. Das Modell wir hier in WINMIRA (von Davier, 1996) so spezifiziert, das eine Klasse ein PCM, die 2. Klasse eine homogene latente Klasse repräsentiert. Letztere enthält dann nur einen einzigen inzidentellen (Personen-) Parameter. Das 2- Klassen Hybridmodell (Rost & von Davier, 1995; von Davier & Rost, 1995; von Davier & Yamamoto, 2007; Yamamoto, 1989) dient hier vor allem zur Bestimmung der Größe der Subpopulation für die ein PCM tatsächlich gilt. Darüber hinaus aber auch zur Bestimmung der prädiktiven- bzw. Konstruktvalidität. Es wird angenommen, dass die 1. Klasse des Hybridmodells auch die Homogenität mit Rücksicht auf Erklärungszusammenhänge erhöht. Zur Beurteilung der Modellgüte dient der parametrische Bootstrap mit 1000 Replikationen (von Davier, 1997) und die Entropie der 2-Klassenlösung. Dargestellt wird die Verteilung vorteilhaft durch ein klassenspezifisches Histogramm, welches idealiter jeweils nur einen Balken bei der Wahrscheinlichkeit von 1 aufweist; nur in diesem Fall ist die Modale mit der probabilistischen Zuordnung der Befragten zu den latenten Klassen identisch. Zur Beurteilung des Itemfits wird sowohl für die 1-Klassenlösung, wie auch für das 2-Klassen Hybridmodell zusätzlich der Q-Index verwendet (Rost & von Davier, 1994; Tarnai & Rost, 1990).

Literatur

Agresti, A (1989). Tutorial on Modeling Ordered Categorical Response Data. Psychological Bulletin, **105**[2], 290-301.

Agresti, A (2002). Categorical Data Analysis. 2nd ed. Hoboken, New Jersey: Wiley.

Andrich, D (1978). Application of a Psychometric Rating Model to Ordered Categories Which are Scored with Successive Integers. Applied Psychological Measurement, **2**[4], 581-594.

Andrich, D (1982). An Extension of the Rasch Model for Ratings Providing Both Location and Dispersion Parameters. Psychometrika, **47**, 105-113.

Bond, TG & CM Fox (2007). Applying the Rasch model. 2nd ed. L. Erlbaum Mahwah, NJ.

De Boeck, P (2008). Random Item IRT Models. Psychometrika, **73**[4], 533-559.

De Boeck, P & M Wilson (2004). Explanatory Item Response Models: A General Linear and Nonlinear Approach. Edited by Paul De Boeck and Mark Wilson . Statistics for Social Science and Public Policy. Ed.: Fienberg, Stephen and Van der Linden, Wim.New York, Berlin: Springer.

Engelhard, G, Jr. (2009). Using Item Response Theory and Model--Data Fit to Conceptualize Differential Item and Person Functioning for Students With Disabilities. Educational and Psychological Measurement, **69**[4], 585-602.

Eun, SK & Myeongsun, Y (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. Structural Equation Modeling, **18**[2], 212-228.

Hatzinger, R & Mair, P (2007). eRm: ein Open Source Paket für IRT Modelle. Department für Statistik und Mathematik Wirtschaftsuniversität Wien .

Jannsen, R, J Scheper & D Peres. (2004). Models with Item and Item Group Predictors. In P. De Boeck & M. Wilson (Eds.) Explanatory Item Response Models: A General Linear and Nonlinear Approach (pp. 189-212). New York, Berlin: Springer.

Linacre, JM & Wright, BD (1994). Reasonable mean square fit values. Rasch Measurement Transactions, **83**[3], 370.

Mair, P & Hatzinger, R (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. Journal of Statistical Software, **20**[9], 1-20.

Masters, GN (1982). A Rasch Model for Partial Credit Scoring. Psychometrika, **47**[2], 149-174.

Masters, GN (1988). The Analysis of Partial Credit Scoring. Applied Measurement in Education, **1**[4], 279-297.

Meulders, M & Y Xie. (2004). Person-by-Item Predictors. In P. De Boeck & M. Wilson (Eds.) Explanatory Item Response Models: A General Linear and Nonlinear Approach (pp. 213-246). New York, Berlin: Springer.

Rabe-Hesketh, S, Skrondal, A & Pickels, A (2004). Generalized Multilevel Structural-Equation Modeling. Psychometrika, **69**[2], 167-190.

Rabe-Hesketh, S, A Skrondal & A Pickles (2004). "GLLAMM Manual". Vol. Working Paper 160 http://www.bepress.com/ucbbiostat/paper160 . U.C. Berkeley Division of Biostatistics Working Paper Series.

Reckase, MD (2009). Multidimensional Item Response Theory. Statistics for Social and Behavioral Sciences. Ed.: Fienberg, Stephen and van der Linden, Wim J.Heidelberg: Springer.

Reeve, BB, Hays, RD, Chih-Hung, C & Perfetto, EM (2007). Applying item response theory to enhance health outcomes assessment. Quality of Life Research, **16**[S1], 1-3. Springer Science & Business Media B.V.

Rijmen, F, Tuerlinckx, F, De Boeck, P & Kuppens P. (2003). A Nonlinear Mixed Model Framework for Item Response Theory. Psychological Methods, **8**[2], 185-205.

Rost, J (1990). Rasch Models In Latent Classes: An Integration of Two Approaches To Item Analysis. Applied Psychological Measurement, **14**[3], 271-282.

Rost, J (1991). A Logistic Mixture Distribution Model for Polychotomous Item Responses. The British Journal for Mathematical and Statistical Psychology, **44**, 75-92.

Rost, J (2004). Lehrbuch Testtheorie- Testkonstruktion. 2nd ed. Bern: Verlag Hans Huber.

Rost, J & von Davier, M (1994). A Conditional Item-Fit Index for Rasch Models. Applied Psychological Measurement, **18**[2], 171-182.

Rost, J & M von Davier. (1995). Mixture Distribution Rasch Models. In G. H. Fischer & I. W. Molenaar (Eds.) Rasch Models: Foundations, Recent Developments and Applications (pp. 257-268). New York, Heidelberg: Springer- Verlag.

Sass, DA & Schmitt, TA (2010). A Comparative Investigation of Rotation Criteria Within Exploratory Factor Analysis. Multivariate Behavioral Research, **45**[1], 73-103.

Skrondal, A & S Rabe-Hesketh (2004). Generalized Latent Variable Modeling; Multilevel, Longitudinal, and Structural Equation Models. Interdisciplinary Statistics Series. Ed.: Keiding, N., Morgan, B., Speed, T., and Van der Heijden, Peter G. M.London, New York: Chapman & Hall/CRC.

Stata: Release 11. Statistical Software, College Station , TX.

Tarnai, C & Rost, J (1990). Identifying Aberrant Response Patterns in the Rasch Model- The Q- Index. Soz.Wiss.Forschungsdokumentation.

Teresi, J & Fleishman, J (2007). Differential Item Functioning and Health Assessment. Quality of Life Research, **16**[S1], 33-42.

von Davier, M (1996). WINMIRA; A Program System for Analysis with the Rasch Model, with the Latent Class Analysis and with the Mixed Rasch Model. Vol. Version 1.74. Kiel: IPN.

von Davier, M (1997). Methoden zur Prüfung probabilistischer Testmodelle. Vol. 157. Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

von Davier, M & J Rost. (1995). Polytomous Mixed Rasch Models. In G. Fischer & I. W. Molenaar (Eds.) Rasch Models. Foundations, Recent Developments, and Applications (pp. 371-382). New York Berlin Heidelberg uaO: Springer Verlag.

von Davier, M & K Yamamoto. (2007). Mixture- distribution models and HYBRID Rasch models. In M. von Davier & C. Carstensen (Eds.) Multivariate and Mixture Distribution Rasch Models. Extensions and Applications (pp. 99-118). New York: Springer-Verlag.

Wilson, M (2005). Constructing Measures: An Item Response Modeling Approach. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilson, M & P De Boeck. (2004). Descriptive and Explanatory Item Response Models. In P. De Boeck & M. Wilson (Eds.) Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach (pp. 43-74). New York: Springer.

Wright, BD & GN Masters (1982). Rating Scale Analysis. Chicago: MESA Press.

Yamamoto, K (1989). HYBRID Model of IRT and Latent Class Models. ETS research report series (RR-89-41). Ed.: Princton NJ: Educational Testing Service.

Zheng, X & Rabe-Hesketh, S (2007). Estimating Parameters of Dichotomous and Ordinal Item Response Models using GLLAMM. The Stata Journal, **7**[3], 313-333.

Zumbo, BD (2007). Three Generations of DIF Analyses: Considering Where It Has Beeen; Where It Is Now, and Where It Is Going. Language Assessment Quarterly, **4**[2], 223-233.